Preparing microdata for public use: Confidentialization and disclosure control

Derek Burk, Senior Research Scientist
Dan Ehrlich, Senior Data Analyst
Lara Cleveland, Director & Prin. Res. Scientist
IPUMS, University of Minnesota

IPUMS.ORG

1

Confidentialization and disclosure control: Outline

Intro
Risk and Utility Assessment
Types of Data Treatments
IPUMS Treatments: Sampling, Suppression, Swapping
Codeshare Demo

IPUMS.ORG

2

Confidentialization and disclosure control:

Intro

Risk and Utility Assessment

Types of Data Treatments

IPUMS Treatments: Sampling, Suppression, Swapping

Codeshare Demo

IPUMS.ORG

3

THE CHALLENGE OF DATA PROTECTION

DATA UTILITY		DATA SAFETY
Huge research potential Better planning	**Data data everywhere**	Data to combine "disclosive"
Make the data accessible	**Good data stewardship**	Protect privacy
More solution options	**Technology**	Solutions in search of problems

IPUMS

4

CHALLENGES: DATA PROTECTION

FIVE SAFES FRAMEWORK

Safe projects	Is this use of the data appropriate, lawful, ethical and sensible?
Safe people	Can the user be trusted to use data in an appropriate manner?
Safe data	Does the data itself contain sufficient information for a potential confidentiality breach?
Safe settings	Does the access facility limit unauthorized use or mistakes?
Safe output	Is the confidentiality maintained for the outputs by the data management system?

https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/

https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework

 5

5

IPUMS INTERNATIONAL

Confidentialization and disclosure control: What **is** confidentialization

- Confidentialization is the process of balancing utility and privacy

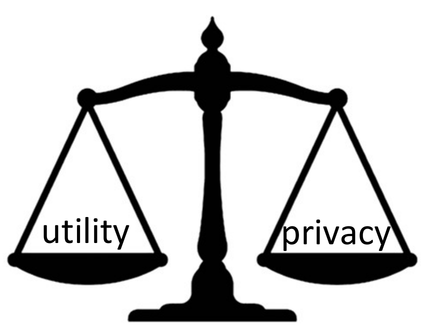
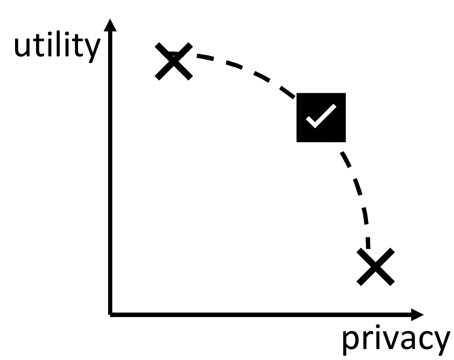


Image: www.needpix.com



IPUMS.ORG

6

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Assessing utility and risk

Utility
Useful ***for what purpose?***

Risk
How might disclosure occur?
What might be disclosed?

IPUMS.ORG

7

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Useful **for what purpose?**

name	address	age	sex	marst	relate
John Doe	123 Main St.	30	male	married	head
Jane Doe	123 Main St.	30	female	married	spouse
Jack Doe	123 Main St.	4	male	*NA*	child
Jill Doe	123 Main St.	2	female	*NA*	child

IPUMS.ORG

8

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Useful **for what purpose?**

name	address	age	sex	marst	relate
		30	male	married	head
		30	female	married	spouse
		4	male	*NA*	child
		2	female	*NA*	child

IPUMS.ORG

9

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Useful **for what purpose?**

id	name	address	age	sex	marst	relate
531	John Doe	123 Main St.	30	male	married	head
532	Jane Doe	123 Main St.	30	female	married	spouse
533	Jack Doe	123 Main St.	4	male	*NA*	child
534	Jill Doe	123 Main St.	2	female	*NA*	child

IPUMS.ORG

10

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Useful **for what purpose?**

id	name	address	age	sex	marst	relate
531			30	male	married	head
532			30	female	married	spouse
533			4	male	*NA*	child
534			2	female	*NA*	child

IPUMS.ORG

11

IPUMS INTERNATIONAL

Confidentialization and disclosure control: **How** might disclosure occur?

region	prov	age	sex	marst	relate	birthplace	occ	educ
3	2	30	male	married	head	San Jose	teacher	Master's
…	…	30	female	married	spouse	…	…	…
…	…	4	male	*NA*	child	…	…	…
…	…	2	female	*NA*	child	…	…	…

"KEY"

IPUMS.ORG

Source: Dupriez and Boyko 2010, page 32

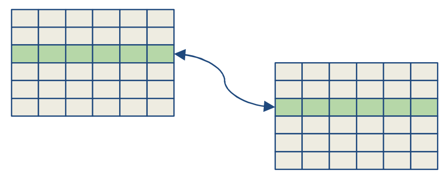
12

Confidentialization and disclosure control: **How** might disclosure occur?

The “nosy neighbor” scenario

Image: www.vectorstock.com

The external archive scenario



Source: Dupriez and Boyko 2010, page 33

IPUMS.ORG

13

Confidentialization and disclosure control: **What** might be disclosed?

region	prov	age	sex	marst	relate	birthplace	occ	educ	income
3	2	30	male	married	head	San Jose	teacher	Master’s	$40,000
…	…	30	female	married	spouse	…	…	…	…
…	…	4	male	*NA*	child	…	…	…	…
…	…	2	female	*NA*	child	…	…	…	…

IPUMS.ORG

14

IPUMS INTERNATIONAL

Confidentialization and disclosure control:

Intro

Risk and Utility Assessment

Types of Data Treatments

IPUMS Treatments: Sampling, Suppression, Swapping

Codeshare Demo

IPUMS.ORG

15

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Risk and Utility Assessment

Apply treatments:

Make data safer

Retain utility

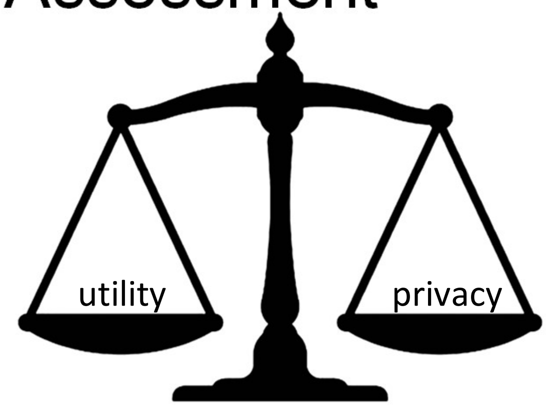


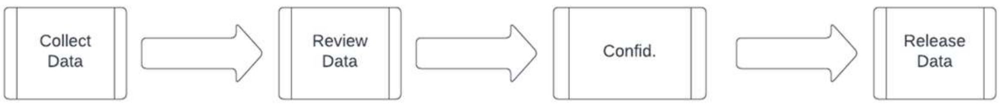
Image: www.needpix.com

IPUMS.ORG

16

Confidentialization and disclosure control: Risk and Utility Assessment

- Within data production framework



17

Confidentialization and disclosure control: Risk Assessment

- Highly Context dependent
- Can occur at different levels, at different times in the project
- Can be informal or formal assessments

18

Confidentialization and disclosure control: Risk Assessment

- Highly Context dependent
- Can occur at different levels, at different times in data life cycle
- Can be informal or formal assessments

IPUMS.ORG

19

Confidentialization and disclosure control: Risk Assessment

- Informal
  - Is this data likely to be targeted?
  - Does this data contain any sensitive questions?
  - Sensitive responses (EG, ethnic minorities)
  - Do certain combinations of variables pose a risk
    - EG, ethnic minority in a specific geographic unit

IPUMS.ORG

20

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Risk Assessment

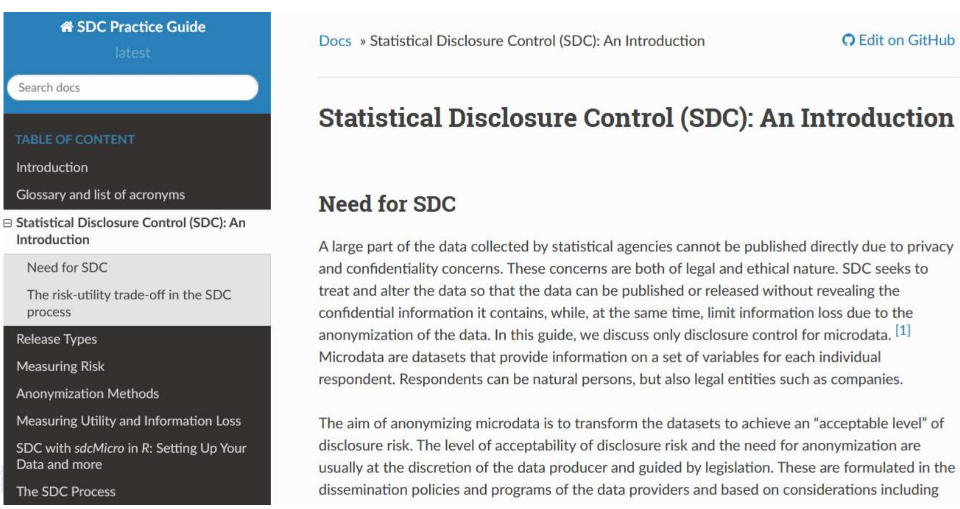
- Formal
  - *k-anonymity, l-diversity, t-closeness; in general:*
    - These methods all center around detecting unique records and how easy it is to individuate records.
  - In practice, these metrics can be complicated to calculate and easily skewed by large number of responses and/or variables.
    - Do not detect noise/confusion or other
    - Does a unique record in a sample represent the same risk as a unique records in the population?

IPUMS.ORG

21

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Risk Assessment


SDC Practice Guide
latest
Search docs
TABLE OF CONTENT
Introduction
Glossary and list of acronyms
Statistical Disclosure Control (SDC): An Introduction
Need for SDC
The risk-utility trade-off in the SDC process
Release Types
Measuring Risk
Anonymization Methods
Measuring Utility and Information Loss
SDC with *sdcMicro* in *R*: Setting Up Your Data and more
The SDC Process

Docs » Statistical Disclosure Control (SDC): An Introduction — Edit on GitHub

Statistical Disclosure Control (SDC): An Introduction

Need for SDC

A large part of the data collected by statistical agencies cannot be published directly due to privacy and confidentiality concerns. These concerns are both of legal and ethical nature. SDC seeks to treat and alter the data so that the data can be published or released without revealing the confidential information it contains, while, at the same time, limit information loss due to the anonymization of the data. In this guide, we discuss only disclosure control for microdata. [1] Microdata are datasets that provide information on a set of variables for each individual respondent. Respondents can be natural persons, but also legal entities such as companies.

The aim of anonymizing microdata is to transform the datasets to achieve an "acceptable level" of disclosure risk. The level of acceptability of disclosure risk and the need for anonymization are usually at the discretion of the data producer and guided by legislation. These are formulated in the dissemination policies and programs of the data providers and based on considerations including

SDC Micro

Risk assessment tool

IPUMS.ORG

22

Confidentialization and disclosure control: Utility Assessment

- Narrow utility
  - Easy for small handful of tests, but unreasonable to know EVERY analysis the public might do
  - Useful to spot-check common cross-tabs (EG: SEX by Geog)
- Broad Utility
  - Not trying to match a specific metric, more concerned with the structure of dataset as a whole
  - a. Can get complicated as many data treatments alter class of the data

IPUMS.ORG

23

CHALLENGES: DATA PROTECTION

FIVE SAFES FRAMEWORK

Safe projects	Is this use of the data appropriate, lawful, ethical and sensible?
Safe people	Can the user be trusted to use data in an appropriate manner?
Safe data	Does the data itself contain sufficient information for a potential confidentiality breach?
Safe settings	Does the access facility limit unauthorized use or mistakes?
Safe output	Is the confidentiality maintained for the outputs by the data management system?

https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/

https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework

 24

24

CHALLENGES:
DATA PROTECTION

DATA SAFETY

DATA UTILITY

FIVE SAFES FRAMEWORK

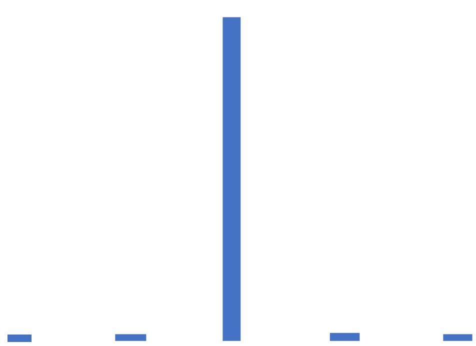
Projects – People – Data – Settings – Output

CHALLENGES:
DATA PROTECTION

DATA SAFETY

DATA UTILITY

- FIVE SAFES FRAMEWORK



- Projects – People – Data – Settings – Outputs

Very safe

Not safe or unknown

Toy data model
Highly treated
Limited detail
No identifiers
Fully public

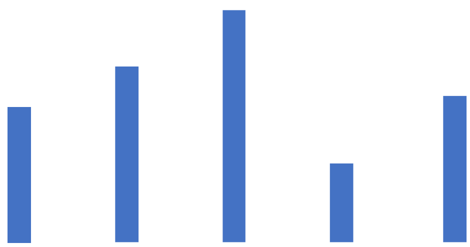
Safe, low utility
or even erroneous

E.g., small toy or
practice datasets

IPUMS

CHALLENGES:
DATA PROTECTION

- **FIVE SAFES FRAMEWORK**



Very safe

Not safe
or unknown

- Projects – People – Data – Settings – Outputs

Scientific Use File
Sampled & lightly treated
Some detail omitted
No identifiers
Screen & vet users

Good analytical power,
some usage barriers

E.g., **IPUMS!** Also DHS, MICS, most surveys

IPUMS

27

CHALLENGES:
DATA PROTECTION

- **FIVE SAFES FRAMEWORK**

Very safe

Not safe
or unknown

- Projects – People – Data – Settings – Outputs

Restricted centers
Little or no infused error
Lots of detail & full files
May have identifiers
Detailed user application

Great analytical power,
limited usage, costly

E.g., Restricted access data centers

IPUMS

28

CHALLENGES:
DATA PROTECTION

DATA SAFETY

DATA UTILITY

- FIVE SAFES FRAMEWORK

- Projects – People – Data – Settings – Outputs

Controlled output systems
Constrained flexibility
Controlled detail
For predictable outputs
Public or internal users

Retain controlled detail, limited flexibility

E.g., Tabulators or in-house admin data query systems

IPUMS

29

CHALLENGES:
DATA PROTECTION

DATA SAFETY

DATA UTILITY

FIVE SAFES FRAMEWORK

Projects – People – Data – Settings – Output

Extra safe data	Safe data-projects-people	Safe people & settings	Controlled output systems
Highly treated Limited detail No identifiers Fully public	Sampled & lightly treated Some detail omitted No identifiers Screen & vet users	Little or no infused error Lots of detail & full files May have identifiers Detailed user application	Constrained flexibility Controlled detail For predictable outputs Public or internal users
Safe, low utility or even erroneous	*Good analytical power, some usage barriers*	*Great analytical power, limited usage, costly*	*Retain controlled detail, limited flexibility*
E.g., small toy or practice datasets	E.g., **IPUMS!** Also DHS, MICS, most surveys	E.g., Restricted access data centers	E.g., Tabulators or in-house admin data query systems

 30

30

IPUMS INTERNATIONAL

Confidentialization and disclosure control:

Intro

Risk and Utility Assessment

Types of Data Treatments

IPUMS Treatments: Sampling, Suppression, Swapping

Codeshare Demo

IPUMS.ORG

31

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Treatments

ST/ESA/STAT/SER.M/67/Rev.3

Department of Economic and Social Affairs
Statistics Division

Principles and Recommendations for Population and Housing Censuses

Revision 3

3.335. As presented in this subsection, there are methods (such as sampling, introduction of random disturbances, recoding and aggregation) that can be used to make such microdata available while still protecting individuals' rights to privacy. All have in common the fact that they sacrifice some information in order to eliminate or greatly reduce the risk of disclosure. However, it is important that census organizations interested in disseminating microdata to outside users should take the appropriate precautions to protect privacy and confidentiality.

IPUMS.ORG

32

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Data Treatments

- Purpose: To modulate risk and utility; **to add uncertainty to data**
- In general datasets with:
    - **Many records and few variables will have inherently low risk due to the small chance of individuation of records.**
    - **Conversely, many variables and few records will result in a high number of unique cases - a potential risk.**
- In general data treatments:
    - **Remove information**: to limit risk but often also limit utility
    - **Add information:** Some treatments add noise/confusion, lowering risk while maintaining utility.

IPUMS.ORG

33

IPUMS INTERNATIONAL

Confidentialization and disclosure control: Treatments

Sampling
Introduction of random disturbances (noise)
    Swapping
    Shuffling
    Perturbing
Suppression
    Recoding and aggregation

IPUMS.ORG

34

IPUMS INTERNATIONAL

Confidentialization and disclosure control:

Intro

Risk and Utility Assessment

Types of Data Treatments

IPUMS Treatments: Sampling, Suppression, Swapping

Codeshare Demo

IPUMS.ORG

35

IPUMS INTERNATIONAL

Sampling

- All modern IPUMS International datasets are samples
- Samples drawn by NSO or IPUMS
- IPUMS provides systematic 1-in-10 sample when possible

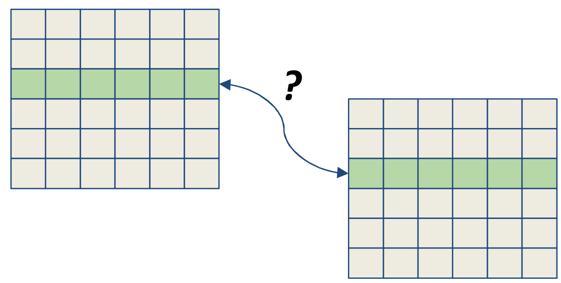
IPUMS.ORG

36

Sampling as disclosure control

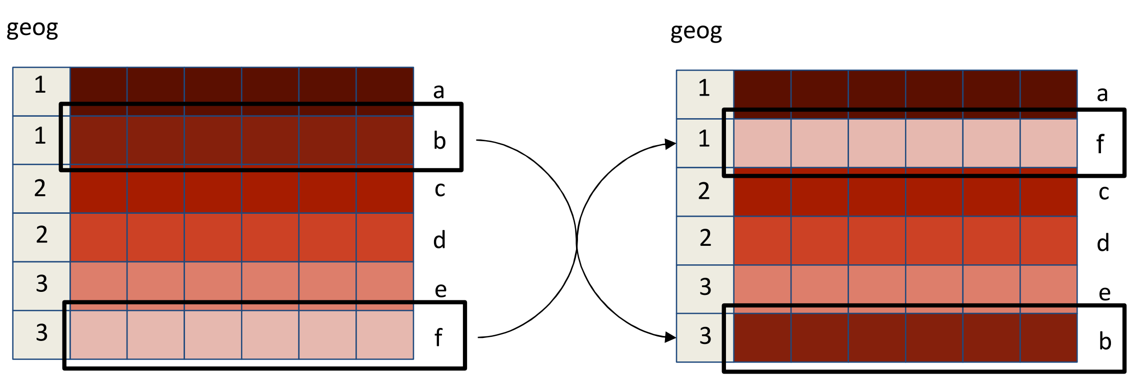
Sampling creates uncertainty:

Is a *sample* unique a *population* unique?



37

Random disturbance: Swapping

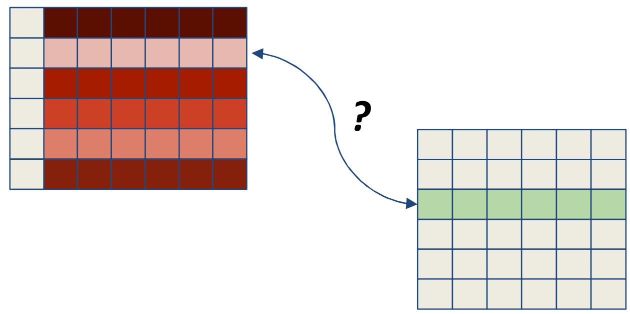


38

Swapping as disclosure control

Swapping creates uncertainty:

Is geography accurate for this particular record?



39

Suppression

occupation	count
…	…
Statisticians	51
Mathematicians	18
Chemists	33
…	…

occupation	count
…	…
Statisticians and mathematicians	69
Chemists	33
…	…

40

IPUMS INTERNATIONAL

Suppression: Top/bottom coding

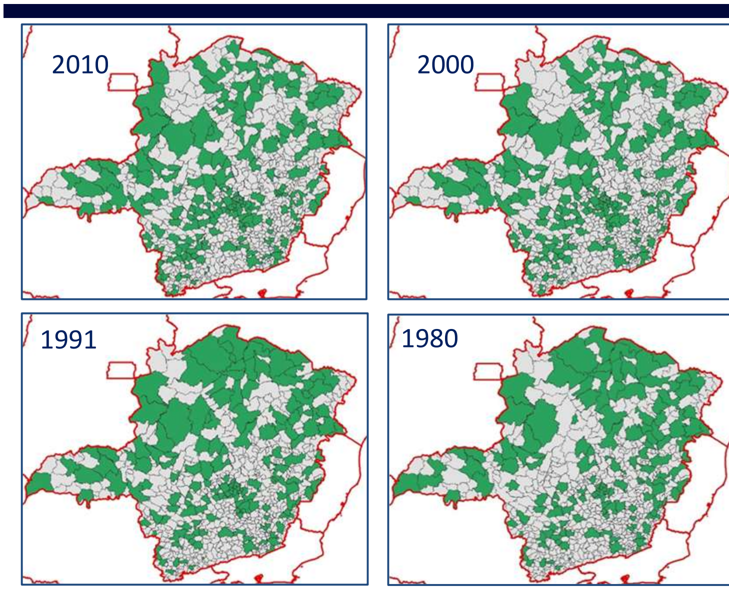
number_of_rooms	count
…	…
20	47
21	21
22	26
23	9

number_of_rooms	count
…	…
20+	103

IPUMS.ORG

41

Regional Confidentialization



Brazil
1980 - 2010

Minas Gerais over the years

< 20,000 people	
2010	673 (79%)
2000	688 (81%)
1991	580 (78%)
1980	565 (80%)

< 20,000
> 20,000

IPUMS INTERNATIONAL

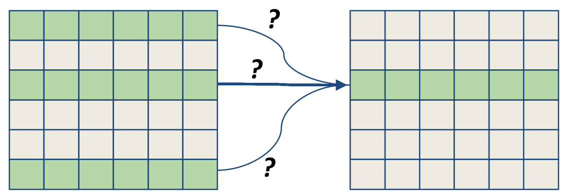
42

IPUMS INTERNATIONAL

Suppression as disclosure control

Suppression reduces uniqueness:

Sample uniques are rarer



IPUMS.ORG

43

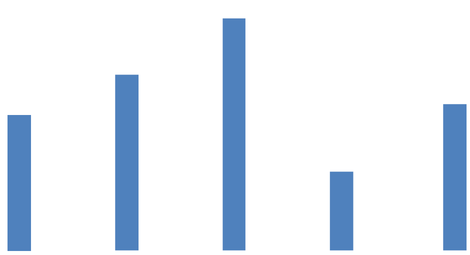
IPUMS INTERNATIONAL

CHALLENGES: DATA PROTECTION

DATA SAFETY

DATA UTILITY

• FIVE SAFES FRAMEWORK



Very safe

Not safe or unknown

• Projects – People – Data – Settings – Outputs

Scientific Use File
Sampled & lightly treated
Some detail omitted
No identifiers
Screen & vet users

Good analytical power, some usage barriers

E.g., **IPUMS!** Also DHS, MICS, most surveys

IPUMS.ORG

IPUMS

44

IPUMS INTERNATIONAL

CHALLENGES:
DATA PROTECTION

FIVE SAFES FRAMEWORK

Safe projects	Is this use of the data appropriate, lawful, ethical and sensible?
Safe people	Can the user be trusted to use data in an appropriate manner?
Safe data	Does the data itself contain sufficient information for a potential confidentiality breach?
Safe settings	Does the access facility limit unauthorized use or mistakes?
Safe output	Is the confidentiality maintained for the outputs by the data management system?

https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/
https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework

IPUMS.ORG

 45

45

IPUMS INTERNATIONAL

Confidentialization and disclosure control:

Intro
Risk and Utility Assessment
Types of Data Treatments
IPUMS-Specific Data Treatments
Codeshare Demo

IPUMS.ORG

46

IPUMS INTERNATIONAL

Inspect Data

```
> ex_data
# A tibble: 26,624 × 19
   SERIAL new_geo0 new_geo1 new_geo2 new_geoFULL URBAN PERNUM  nper RELATE             AGE      SEX      MARST      BIRTHYR    BIRTHMO   CITIZEN NATION        EDATTAIN
    <dbl> <chr>       <dbl>    <dbl> <chr>       <chr>  <dbl> <dbl> <dbl+lbl>          <dbl+l>  <dbl+l>  <dbl+lbl>  <dbl+lbl>  <dbl+lb>  <dbl+l> <dbl+lbl>     <dbl+lbl>
 1      1 01              1        1 0111        RURAL      1     6 1 [HEAD]           45 [45]  1 [Mal… 210 [Mar… 1955 [195…  4 [Apr… 1 [Cit… 13040 [Mor… 120 [Som…
 2      1 01              1        1 0111        RURAL      2     6 2 [SPOUSE/PARTN… 48 [48]  2 [Fem… 210 [Mar… 1952 [195…  8 [Aug… 1 [Cit… 13040 [Mor… 110 [No …
 3      1 01              1        1 0111        RURAL      3     6 3 [CHILD]          18 [18]  2 [Fem… 100 [SIN… 1982 [198…  1 [Jan… 1 [Cit… 13040 [Mor… 221 [Gen…
 4      1 01              1        1 0111        RURAL      4     6 3 [CHILD]          16 [16]  1 [Mal… 100 [SIN… 1984 [198…  5 [May] 1 [Cit… 13040 [Mor… 212 [Pri…
 5      1 01              1        1 0111        RURAL      5     6 3 [CHILD]          14 [14]  2 [Fem… 100 [SIN… 1986 [198…  5 [May] 1 [Cit… 13040 [Mor… 110 [No …
 6      1 01              1        1 0111        RURAL      6     6 3 [CHILD]          12 [12]  2 [Fem… 100 [SIN… 1988 [198… 12 [Dec… 1 [Cit… 13040 [Mor… 120 [Som…
 7      2 01              1        1 0111        RURAL      1     8 1 [HEAD]           48 [48]  1 [Mal… 210 [Mar… 1952 [195…  7 [Jul… 1 [Cit… 13040 [Mor… 312 [Som…
 8      2 01              1        1 0111        RURAL      2     8 2 [SPOUSE/PARTN… 43 [43]  2 [Fem… 210 [Mar… 1957 [195…  2 [Feb… 1 [Cit… 13040 [Mor… 110 [No …
 9      2 01              1        1 0111        RURAL      3     8 3 [CHILD]          17 [17]  1 [Mal… 100 [SIN… 1983 [198… 11 [Nov… 1 [Cit… 13040 [Mor… 311 [Gen…
10      2 01              1        1 0111        RURAL      4     8 3 [CHILD]          15 [15]  1 [Mal… 100 [SIN… 1985 [198…  9 [Sep… 1 [Cit… 13040 [Mor… 221 [Gen…
# ℹ 26,614 more rows
# ℹ 2 more variables: OCC <dbl+lbl>, DISABLED <dbl+lbl>
```

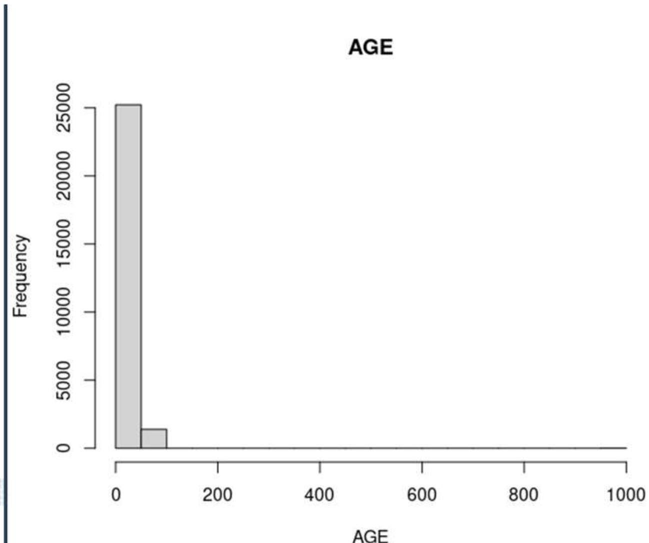
IPUMS.ORG

IPUMS INTERNATIONAL

Inspect Variables

```
Labels:
 value              label
     0   Less than 1 year
     1             1 year
     2            2 years
     3                  3
     4                  4
     5                  5
     6                  6
     7                  7
     8                  8
     9                  9
```

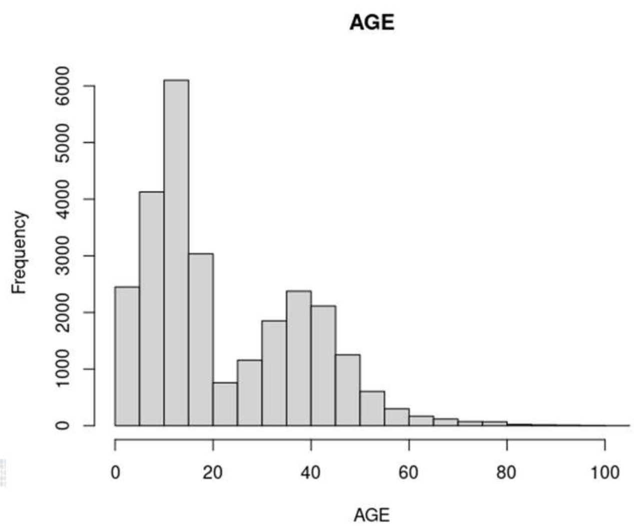
```
    93                 93
    94                 94
    95                 95
    96                 96
    97                 97
    98                 98
    99                 99
   100               100+
   999 Not reported/missing
```



Inspect Variables - Recode Special Cases

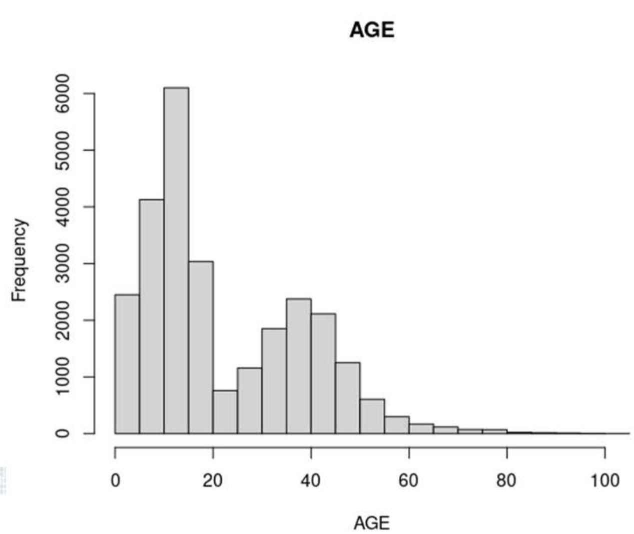
- Recode Special Values
  - AGE

  For cases when AGE ==999, set to NA
- ex_data <- ex_data %>% mutate(AGE2 = if_else(AGE==999, NA_integer_, AGE))



Top-coding Age

- Skewed distributions of age can mean low-representation.
- Consider top-coding and/or grouping into 5-yr age cohorts
  - Raises utility for analysis
- IPUMS releases both a top-coded integer-age as well as a 5-yr age-group



IPUMS INTERNATIONAL

Consider Related Variables

- Age: more than just age
  - If you top-code age, make sure to bottom-code Birth Year
- Other commonly related variables:
  - Country of birth/nationality
  - Occupation, industry, etc

IPUMS.ORG

51

IPUMS INTERNATIONAL

Detecting and Recoding of Small Cells

- Occupation has over 100 levels of responses, some highly represented, some very minimally represented.

IPUMS.ORG

52

IPUMS INTERNATIONAL

Recoding of Nationality/Citizenship

IPUMS.ORG

53

IPUMS INTERNATIONAL

References

Dupriez, Olivier and Ernie Boyko. 2010. "Dissemination of Microdata Files. Formulating Policies and Procedures", International Household Survey Network, IHSN Working Paper No 005. http://ihsn.org/sites/default/files/resources/IHSN-WP005.pdf

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Naylor, Eric Schulte Nordholt, Giovanni Seri, and Peter-Paul De Wolf. 2010. *Handbook on Statistical Disclosure Control, version 1.2*. ESSNet SDC. https://cros-legacy.ec.europa.eu/system/files/SDC_Handbook.pdf

IPUMS.ORG

54

IPUMS INTERNATIONAL

References

risk assessment:
sdcMicro User Guide: https://sdcpractice.readthedocs.io/en/latest/intro.html
sdcMicro is an R package produced by The World Bank, in collaboration with Intl Household Survey Network (IHSN), PARIS21 (OECD), Statistics Austria and the Vienna University of Technology.

k-anonymity:
Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" (PDF). Harvard Data Privacy Lab. Retrieved April 12, 2017.

critiques of k-anon and expansions (t-closeness):
N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 2007, pp. 106-115, doi: 10.1109/ICDE.2007.367856.

Utility Assessment: sdcMicro material and the synthpop team (Based at University of Edinburgh):
"Assessing, visualizing and improving the utility of synthetic data." Raab, G. UNECE CONFERENCE OF EUROPEAN STATISTICIANS: Expert Meeting on Statistical Data Confidentiality
1-3 December 2021, Poland
https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Raab_AD.pdf

IPUMS.ORG

55